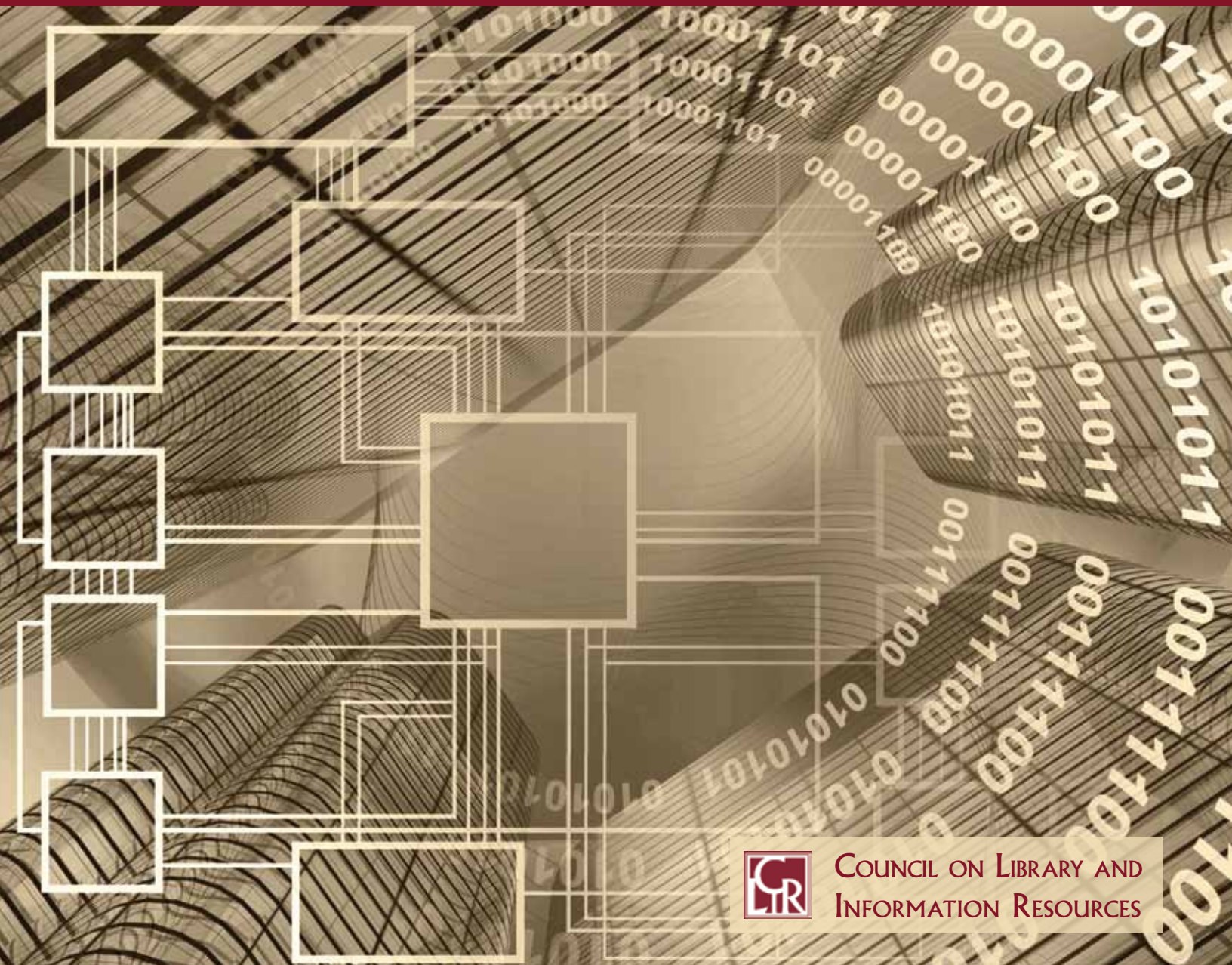
Core Infrastructure Considerations for Large Digital Libraries

by Geneva Henry

July 2012



Council on Library and Information Resources

Core Infrastructure Considerations for Large Digital Libraries

by Geneva Henry

July 2012

Council on Library and Information Resources

DLF
Digital Library Federation
Program of the Council on Library and Information Resources

ISBN 978-1-932326-41-3
CLIR Publication No. 153
Published by:

Council on Library and Information Resources
1707 L Street NW, Suite 650
Washington, DC 20036
Web site at http://www.clir.org

Additional copies are available for $15 each. Orders must be placed through CLIR's Web site.
This publication is also available online at http://www.clir.org/pubs/reports/pub153.

The paper in this publication meets the minimum requirements of the American National Standard for Information Sciences—Permanence of Paper for Printed Library Materials ANSI Z39.48-1984.

Cover photo: © shutterstock.com/archerix

Contents

About the Author

Geneva Henry is the executive director of Digital Scholarship Services at Rice University's Fondren Library. She joined Rice in 2000 to start Rice's digital library initiative, which has grown to include many projects, both grant funded and internally sponsored. In addition, Henry is currently a Research Library Leadership Fellow with the Association of Research Libraries. In 2006, she served as a distinguished fellow for the Digital Library Federation, where she was part of the Abstract Services Framework group working to develop a framework of digital library services. From March 2002 through June 2005, she was the executive director for the Connexions open textbook project, helping to shape and launch that project. She is active in professional organizations and conferences related to digital libraries, and she serves as a board member for several organizations and projects, including CLOCKSS, the Digital Scholarship Commons (DiSC), and the European Union's DL.org project. Prior to joining Rice, Henry was a senior information technology architect and program manager with IBM, where she was involved in several complex systems programs for government agencies, universities, and museums worldwide.

Acknowledgments

This report was funded by a grant to the Council on Library and Information Resources (CLIR) from The Andrew W. Mellon Foundation for developing a prototype for the Digital Public Library of America. Richard Urban provided research support for metadata approaches and digital library architectures. Rachel Frick from the Digital Library Federation assisted in the review and management for the overall grant of which this is a part. The team from the Center for Informatics Research in Science and Scholarship (CIRSS) in the Graduate School of Library and Information Science at the University of Illinois, Urbana-Champaign, provided many valuable insights through their Digital Public Library of America (DPLA) beta sprint efforts and review of drafts of this report. Numerous reviewers of the preliminary draft suggested enhancements that have been incorporated into the final version. Finally, Kathlin Smith's diligent editing of the text improved the final report by making it more coherent.

1. Introduction

Digital library initiatives over the past 15 years have made a wealth of scholarly materials available online. The ready accessibility of these resources has resulted in new business models for information delivery and a cultural shift in the way that information is distributed. Mass digitization initiatives that have focused on digitizing scholarly print materials from leading research libraries have been key to creating much of the digital scholarship that is available.

Mass digitization refers to scanning efforts that do not select particular materials for digitization, but rather convert materials to a digitized format on an "industrial scale," digitizing at high speed and applying optical character recognition (OCR) technologies to the texts to make them searchable (Coyle 2006). The goal of mass digitization is to digitize everything and to minimize human intervention. Examples of libraries created through mass digitization include Google Books, HathiTrust, and The Open Content Alliance from the Internet Archive.

Coyle distinguishes mass digitization efforts from large-scale digital projects. She describes the latter as more selective in the resources that they digitize and include in their collections; in addition, large-scale digital projects may comprise distributed collections to create a larger, more comprehensive digital library than any of the contributing collections alone could provide. Examples of large-scale digital projects that offer interesting resource management approaches include Europeana; the National Science Digital Library (NSDL); the Institute of Museum and Library Services Digital Collections and Content (IMLS DCC)/Opening History; the California Digital Library (CDL); Networked Infrastructure for Nineteenth-Century Electronic Scholarship (NINES); and some sample U.S. statewide digital libraries, such as Florida Memory, Digital Library of Georgia, Digital Commonwealth (Massachusetts), and Seeking Michigan.

The purpose of this study is to examine the core infrastructure elements of systems that manage large quantities of digital materials that one would think of as a digital library, whether created through mass digitization efforts or through large-scale digital projects. An examination of the infrastructures of a few make it possible to understand the diverse approaches each has taken to manage digital content. Although there are many smaller, specialized digital libraries, creating libraries through mass digitization and large-scale efforts poses particular challenges that come with scale and breadth of

materials. This study focuses on large, noncommercial digital libraries, as the infrastructures of large commercial collections are not generally open for analysis.

The study is guided by the following questions related to the core digital infrastructures used for managing content:

- *Storage and Content Delivery*: What are the storage approaches? What kinds of servers are needed? What databases and repository platforms are part of the infrastructure? How is content distributed? What are the assumptions and requirements for content formats?
- *Metadata Approaches and Harvesting*: What metadata formats are supported or required? How are content and its associated metadata managed (e.g., separated, tightly coupled)? What do the metadata look like (e.g., standards used, markup, mandatory fields)? How are metadata harvested?
- *Search and Discovery*: What discovery approaches are used to find content in the collections (e.g., full-text search, image search, faceted browsing, type of search engine used)?
- *Services and Applications*: Are special applications required to use the content (e.g., page turning, jpeg2000 viewers, etc.)? What services are provided?
- *System Sustainability*: What are the system sustainability practices and policies for the digital library?

This study focuses on understanding different approaches to managing large digital libraries; it is not a comprehensive review and comparison of all the systems surveyed.

2. Storage and Content Delivery

Managing large amounts of digital information requires a robust server and a storage system that is reliable and has strong performance. While small digital collections do not need to worry as much about the details of storage, large digital libraries are likely to see significantly high levels of activity from users who will expect to access resources quickly and reliably. Decisions about the storage hardware, server allocations, databases, and distribution approaches, along with bandwidth considerations, are key in establishing the digital library as a resource that teachers, students, researchers, and the general public regard as reliable. If a user cannot access content easily, if a video can be viewed only after it has been downloaded, if music does not play at an even speed, or if content becomes corrupted or lost over time, for example, the community will not view the digital library as a credible source of information.

2.1 Storage

Storage decisions for large digital libraries have the most impact when the content is centrally located rather than distributed among systems that have differing architectures for managing the content.

Repositories that are content aggregators increasingly have looked to clustered storage solutions to provide reliable and robust performance. Disc speed, failover capabilities, and automatic error correction functionality found in clustered storage solutions help ensure high availability of the content and a performance that is adequate to meet user needs. Depending on the type of content in the digital library and the approach adopted for delivering it, higher speed storage may be needed to ensure adequate performance (e.g., for streaming music); text-only content may not require a speed as high, but will still require failover and automatic error correction.

Commercial "cloud" storage solutions have become increasingly popular, but not generally for the large-scale digital libraries reviewed in this study. Cloud storage can be very expensive under the current pricing models, because the amount of output, or use, of the data once they are in the cloud drives the primary cost. If users are downloading or streaming significant amounts of content, the cost of using current cloud storage vendors that provide a robust and reliable storage configuration may not be feasible. For example, the current pricing structure for Amazon's S3 service is tiered, with the first year of storage free for 5 GB of Amazon S3 storage, 20,000 Get requests, 2,000 Put requests, and 15 GB of data transfer out each month (Amazon Web Services, LLC n.d.). Storage costs are very reasonable, but data transfers out of the cloud for an active site can be cost prohibitive. The data transfer pricing model, as of June 1, 2012, is shown in table 1. Amazon offers redundancy storage of the content for a reduced rate, providing a means of creating a backup of the primary data store.

	Pricing
Data Transfer IN	
All data transfer in	$0.000 per GB
Data Transfer OUT	
First 1 GB / month	$0.000 per GB
Up to 10 TB / month	$0.120 per GB
Next 40 TB / month	$0.090 per GB
Next 100 TB / month	$0.070 per GB
Next 350 TB / month	$0.050 per GB
Next 524 TB / month	Contact Amazon
Next 4 PB / month	Contact Amazon
Greater than 5 PB / month	Contact Amazon

Table 1. Data transfer costs for Amazon S3, June 1, 2012 (Amazon Web Services, LLC)

HathiTrust, an example of a centralized, aggregated collection, uses the Isilon clustered storage system, which is highly scalable with the addition of new nodes to the storage cluster (HathiTrust, n.d.). The Isilon's OneFS operating system allows all storage nodes in the cluster to know the full file system layout, permitting each to act as a peer and to service any incoming requests. Redundancy is

controlled at the volume, directory, and file levels, making this storage solution highly reliable (Isilon Systems 2011). Similarly architected storage solutions are available from other storage providers as well. DataDirect Networks, for example, provides high availability and responsive input/output using different storage components and software (DataDirect Networks, Inc. 2012).

Large digital libraries may choose to approach the collections from a federated perspective, relying on the original content providers to store their own content. Metadata are centralized to support resource discovery across the many repositories. When federating hundreds or thousands of collections, reliable and fast performance for searching the digital library is critical and the storage decisions can have a significant impact. The choice of metadata representations can affect the storage solutions selected. Large federated collections, such as Europeana and IMLS DCC/Opening History (Center for Informatics Research in Science and Scholarship 2009), host the metadata for many collections whose items are stored in distributed locations, resulting in a large quantity of data that must be stored in a way that allows them to be searched efficiently (Dekkers, Gradmann, and Molendijk 2011). The structure for the metadata's storage—for example, a large flat set of metadata records vs. relational or Resource Description Framework (RDF) representations—will affect how the search will be performed and whether distributing the metadata across multiple storage servers can be readily supported with adequate performance.

2.2 Servers

Servers are critical in providing a reliable platform for the digital library. Systems that have robust architectures to accommodate high performance under varying usage loads have configured a number of services that distribute the activity load over multiple servers; they run services on different servers and use load balancers to handle activity dynamically so that peak usage periods will not result in a server crash. Server hardware for large and mass digitization projects can vary widely. HathiTrust relies on commodity Intel-based servers that run the Linux operating system. Most large digital libraries do not specify their server hardware or operating systems. Europeana maintains a service level agreement (SLA) with Vancis, "a private company with firm roots in Academia and therefore with excellent and relatively cheap connectivity directly to the Amsterdam Internet Exchange (AMS-IX)" (Dekkers, Gradmann, and Molendijk 2011); those servers also run the Linux operating system.

Redundant configurations and load balancing provide high availability and overall system reliability. The largest digital libraries are conscious of the need to provide this type of robust infrastructure, although the approach for configuring multiple servers varies by project. Europeana runs its Web servers, database servers, and image servers on separate machines. The most commonly used Web server is Apache.

Repositories that support audio and video resources may need to consider streaming servers as part of their server infrastructure. Streaming of multimedia resources can be advantageous, as users will not need to download these often large files prior to listening to or viewing them. Because there are multiple streaming servers available on the market, platform compatibility for end users can be a key consideration. For example, Windows platforms and OSX platforms may have differing requirements for delivery applications. If users are required to have a plug-in before they can stream the multimedia resources, they will need to have sufficient knowledge or technical support for installing the plug-in on their computer. In public settings or in many institutions, users are not given permission to install the applications needed to use online resources. If streaming servers are planned as part of the infrastructure, formatting the content to work with multiple streaming servers would be prudent. Permitting download of the resources, although less efficient, will allow users who cannot stream them to access them for listening or viewing on their computers.

2.3 Databases and Repository Platforms

Decisions about managing the content and metadata often depend on the types of data to be stored and the search approach to be used. Repository platforms such as DSpace and CONTENTdm are popular choices for many U.S. state digital library collections and organizations that do not have a large staff of programmers to support functional system development. An example of a CONTENTdm site is Seeking Michigan (SeekingMichigan.org 2008), the online archive for Michigan's digital cultural heritage resources.

The larger digital libraries reviewed in this study, apart from the state-level collections, have developed their own platforms for managing content. NSDL has developed the NCore platform, which uses the Fedora repository software and provides its own data model as well as a suite of tools to support NSDL (Krafft, Birkland, and Cramer 2008). The project also created MPTStore as a means for robustly managing RDF triples (Cornell University 2006), because existing solutions for managing triple stores did not adequately support very large quantities of RDF data with the scalability, performance, and reliability that NSDL required. SQL relational databases, both commercial and open source, are common back-end databases for managing information in large digital repositories. HathiTrust uses MySQL, which provides a good environment for handling the content and metadata formats used in that repository.

As projects grow and mature, decisions about databases and platforms are likely to be revisited. The Europeana project, for example, has found that the Lucene/Solr search engine provides an adequate database for the metadata it manages. Europeana Semantic Elements (ESE) provides a flat data model that works well using Solr as the database (Dekkers, Gradmann, Meghini, et al. 2011). As more complex approaches are considered for a newer Europeana Data

Model, however, thought should be given to alternative database implementations. Under consideration are

- modifying Solr, such as caching and precomputing search results to accommodate continued growth;
- moving to the noSQL document database and combining that with Solr; and
- developing a management solution for RDF triple stores. This, however, is of some concern because billions of triples would be needed to represent the content. Such a solution could have a significant negative impact on performance.

The choices made regarding content format, metadata, and search and browse strategies, as well as the ability of the project to maintain the technology, should inform decisions about selecting a repository platform or data management infrastructures.

2.4 Content Distribution and Format Assumptions

Decisions about content management may affect the approach taken to ensuring the content is available when needed. One means of providing reliable access is to mirror (i.e., replicate) the digital library in multiple locations so that if there is a problem at one location, the mirrored site can provide continued access to the resources and services. Mirroring across multiple geographic locations can facilitate better, more reliable delivery of the resources than does a single site. Examples of large archives that have mirror sites include the Internet Archive, with the production site in San Francisco and the mirror site at Bibliotheca Alexandrina in Egypt; Europeana, with mirrored sites at its host provider's data centers in Amsterdam and Almere in The Netherlands (it is unclear which site is the primary production site); and HathiTrust, with the production site at the University of Michigan and the mirror site at Indiana University's Indianapolis campus.

Mirroring is not the only means of configuring systems to ensure continued high availability. The projects and organizations mentioned in the previous paragraph are also attentive to load balancing their servers and distributing functions across multiple servers. High availability configurations will generally support scalability so that as traffic increases and content grows, systems will continue to meet user demands. NSDL supports high availability by using a Fedora-level transaction journaling system developed for the project. This system allows for replication of transactions in real time to two "follower" systems, ensuring minimal downtime in the event of updates and failures (Krafft, Birkland, and Cramer 2008).

Backup and restore services are critical for the recovery of content in the event of a catastrophic failure. Such services are among the simplest but most fundamental for any repository. The frequency of backups and the media chosen for backups (e.g., tape, disk) vary across projects. As discussed previously, storage solutions can have redundancy built into their architecture. In addition to the redundancy provided in its clustered storage, HathiTrust provides

system-level backup and restore functionality with both file system backup and database backup through Tivoli Storage Manager software.

The choice between hosting content centrally or federating the content, whereby content providers give continual access to the resources identified by the large digital library, can affect the reliability of the overall system. Reliable configurations of large federated digital libraries will help ensure resource discovery and allow the use of tools and services that the project or organization provides, but the federated configuration means that it will not be able to guarantee access to the digital content whenever a user chooses to view or download it. Unless there are SLAs in place with the content providers that require them to maintain a high availability configuration, the large federated digital library may not be able to fulfill the user requests for access. Europeana, for example, has mirrored sites and highly available, load-balanced servers with distributed functionality, but users will be taken to the content provider's site to view the actual resource. If that site is down, or, worse, if it has suffered a catastrophic failure and has lost content, Europeana will have no way to retrieve the resource.

It is a challenge for large federated systems to ensure that content is reliably available from the contributing content providers, especially when the providers have developed their collections through grant funding and have no commitment to maintain the system or to ensure availability after the grant ends. Organizations may also move their collections without notifying the systems that harvest and aggregate their metadata, resulting in an error message to the user who attempts to access the content. IMLS DCC/Opening History (Center for Informatics Research in Science and Scholarship 2009) maintains a federated system that relies on many grant-funded projects to maintain their content after the grant has ended. Project staff members work to keep the links current, but with the number of collections included, it is a challenge. Digital Commonwealth, the federated repository for the state of Massachusetts, describes itself as a portal, relying on a minimal collection of information centrally and redirecting users to the content provider sites (Digital Commonwealth 2007). This approach minimizes the resources needed to keep Digital Commonwealth running, but carries the risk of digital assets being unavailable for users who wish to access them.

Centralized repositories such as HathiTrust have more control over the content and availability of the overall system, but they face other challenges. These include the ability to grow content rapidly and the willingness of content providers to give their content to the centralized digital library. Concerns about rights, attribution, and the ability to attract users to the provider's own site or facility may make some content providers hesitant to contribute their materials to a centralized large digital library. Central repositories often require content to be provided in specific supported formats. If contributors do not have enough staffing or knowledge to comply with those standards, the burden falls on the large digital library either to do

the work or to decide not to accept the content. Consequently, staffing requirements are generally much higher for centralized digital library systems.

The flexibility of formats that can be accepted for inclusion in large digital libraries affects the participation of content providers. NINES, a large digital library, had initially assumed that the format of contributed content would be transcribed text with XML markup, most notably Text Encoding Initiative (TEI) (nines.org 2012). As repository platforms became more pervasive, the organization realized that it needed greater flexibility in its guidelines regarding both content and metadata formats, so NINES modified its requirements to better accommodate content in repositories such as Fedora, DSpace, and CONTENTdm. This has allowed NINES to scale to a much larger federated collection.

3. Metadata Approaches and Harvesting

Whether centralized or federated, large digital libraries rely on standardized metadata to provide information about the resources in their collections. Federated digital libraries harvest metadata from their contributing collections; some also harvest full digital text to support resource discovery.

3.1 Metadata Formats

Many large digital libraries use the Dublin Core (DC) metadata standard, although most add elements to support the services they provide. Large digital collections that use a common repository platform, such as many of the large U.S. state digital libraries, do less customization. Aggregators who harvest metadata from content providers usually publish metadata requirements, and the providers are often responsible for ensuring that their metadata meet the specifications. Large digital libraries with sufficient staff support may take the metadata from the providers, along with a mapping, and themselves get the metadata into the needed format for the repository.

As the quantity of digital objects grows to hundreds of thousands of items, organizations sometimes find that they need a new approach to their digital library data model. NINES initially planned to manage all content centrally, with protocols for submission; it was assumed that content would be marked up in TEI XML and that a METS wrapper would be used to describe and link the resources. After a few years, however, it became clear that the collections in NINES could grow much more significantly using a federated approach and adopting a more widely used metadata standard that would capture content stored in scholarly repositories. NINES now requires content providers to submit their metadata in a DC "flavor" of RDF (Nowviskie and McGann 2005). Although the burden is on the provider to conform to the metadata requirements, providers can now more easily participate in NINES.

As noted earlier, Europeana initially used ESE to provide

descriptive information about content. ESE is described as "Dublin Core plus a few project-specific elements" (Dekkers, Gradmann, Meghini, et al. 2011). The data are "flat," enabling Solr to serve as the repository for the metadata as well as the search engine. Although this arrangement is convenient, the team would like to realize some improvements. They are moving to the Europeana Data Model (EDM) where contributors supply their metadata with a mapping file that maps it to EDM. During the mapping ingestion process, enhancements that enrich the metadata will occur, such as applying named entity recognition, linking to Geonames or Virtual International Authority File (VIAF) records, and normalizing date values. All enriched and normalized data are stored in separate fields, or aggregations, next to the original record. Implementing these enhancements in an RDF representation will enable support for links to related content, helping to foster a linked open data environment (Heath and Bizer 2011).

Content in HathiTrust is described using a METS file for preservation, structural, and technical metadata. PREMIS preservation metadata are updated whenever actions occur on an object. In addition, MARC21 bibliographic records are held for all of the content. Maintaining traditional cataloging records for digital library content is not common among large digital libraries, though creating bibliographic records from digital metadata would not be difficult.

Metadata in NSDL consist of DC elements plus some additional NSDL qualified elements. Additionally, the project stores aggregation objects and agent objects. Relationships among the objects in the digital library are expressed as RDF triples. Support for nested aggregations enables richer searches and relationships between digital objects.

3.2 Management of Metadata with Content

Metadata can be either loosely or tightly coupled with its corresponding digital content objects. Proponents of loose coupling argue that separating the metadata from the objects supports better scalability for the digital library and better performance for searching and browsing. Since metadata is generally structured and comparatively very small, it can be searched rapidly and efficiently. By contrast, the objects being described by the metadata can be extremely large, especially when they are multimedia objects. Keeping the objects and metadata loosely coupled allows additional storage to be added and objects migrated across that new storage without affecting the metadata, except for noting where the object resides on the disc so it can be retrieved when requested. Searching, therefore, remains robust, delivering better overall performance for end users. Platforms such as DSpace support this type of architecture for metadata.

Advocates for tighter coupling of metadata with the content object note that metadata are essential components of the digital resource and should always accompany it as it moves. There should be no risk associated with losing the metadata because of their separation from the resource they describe. It is also critical for the digital

preservation of the resource to have the metadata bundled with the object. Large digital libraries are starting to employ both approaches: structured metadata, such as DC, loosely coupled with the content; and a descriptive metadata file, such as METS, included with the digital objects.

Both NSDL and Europeana are metadata aggregators, with NSDL also supporting aggregated content. HathiTrust maintains a METS file with the digital content as well as catalog records for each resource in its integrated library system, Aleph. NINES keeps the metadata tightly coupled, with RDF embedded with the content. There is no clear right or wrong approach; each has its advantages and drawbacks. The content, the way in which the metadata will be used, the responsibility for creating the required metadata, and the long-term preservation required are important considerations in decisions about the best approach to metadata for the project.

3.3 Harvesting and Content Ingestion

Harvesting metadata is a way to gather descriptive information about items in distributed collections for a federated digital library. Having the metadata centralized can enable common support for functions such as discovery services, timelines, tag clouds, or geospatial visualization that can be used with all the federated collections, even though the content remains distributed. Approaches to the harvest of metadata differ, but many rely on content providers to make their data available to the large digital library in a known format. Sophisticated digital libraries such as the IMLS DCC/Opening History digital library at the University of Illinois will accept multiple metadata formats. Generally, guidelines are provided that describe the formats required by the harvesters; where the guidelines are more flexible about acceptable formats, staff supporting the federated digital library will receive the harvested metadata and transform it into the format needed for the federated repository.

As large digital libraries have discovered, harvesting metadata from varying collections can be a big challenge. Although guidelines may have been given, content providers may not follow them exactly, necessitating additional work by the metadata aggregator. Large digital libraries that have lean staffing may find it impossible to perform this additional work and may simply exclude the collection. Metadata can be harvested from collections using a DC metadata format by means of a DC harvester such as Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH). The IMLS DCC/Opening History beta sprint for the Digital Public Library of America (DPLA) demonstrates the effectiveness of this approach, but there are also challenges with harvesting metadata from collections that use non-DC formats (e.g., RDF and TEI markup).

Standards such as TEI and RDF are flexible; although flexibility has advantages for describing many facets of a work, it makes standard processing difficult. IMLS DCC/Opening History has successfully shared records with collections that have embedded TEI

headers for the object metadata where there is a clearly defined XML Schema Definition (XSD). Embedding Metadata Object Description Schema (MODS) and Metadata Encoding and Transmission Standard (METS) in metadata has also been successful. As Tim Cole has noted, the challenge with TEI is that there is not a canonical XSD for it. The modularity of TEI allows providers to define an XSD for a given module. There does exist a "tei-all.xsd" to use as a default, but it does not represent the breadth of TEI usage. TEI headers are generally the key elements for exchanging metadata files in collections marked up with TEI, and providers will create their own XSD for the headers. A similar challenge exists with RDF. As with TEI, there is no canonical XSD to support RDF; indeed, there is resistance within the RDF community to the development of an XSD, because it would be almost impossible to capture all of the nuances in the RDF data model (Tim Cole, personal communication). The IMLS DCC/Opening History team has, however, found some workarounds to enable sharing of RDF metadata.

Using a crawling approach to harvest metadata records is one way to overcome variances in formats. In its DPLA beta sprint submission, the CDL used Apache Nutch (Apache Foundation 2011) to crawl sites to form its search index (California Digital Library 2011). This approach can have significant performance problems, especially if it is not configured for a high-performance computing environment, and reconciling metadata fields can be a challenge. It does, however, make it possible to search collections with metadata formats that are less structured than a format like Dublin Core. A combination of harvesting and crawling could be very powerful in gathering metadata for access to a variety of collections.

As noted in section 3.1, with the move to EDM, Europeana has taken on the metadata ingestion work of collecting the metadata and the mapping files from the content providers, then enhancing the metadata when it is ingested. HathiTrust ingests its content through a batch process using the Google Return Object-Oriented Validation Environment (GROOVE). Originally developed for the Google Books project, GROOVE is a means of batch ingesting works from other collections as well. HathiTrust makes its DC metadata available to OAI-PMH harvesters.

4. Search and Discovery

Search technologies have improved significantly over the past decade. For text, large repositories most often use the Lucene search engine, an open source information retrieval platform (The Apache Software Foundation n.d.). Solr is a scalable search engine that uses the Lucene library, and many digital repositories implement it to support full-text search, hit highlighting, faceted search, dynamic clustering, database integration, and rich document (e.g., Word, PDF) handling (The Apache Software Foundation n.d.).

Europeana supports simple and advanced searches, but it searches only metadata, not full text. The Lucene-Solr search engine

has worked well for the ESE data model. The flat catalog list simplifies searching, but the move to EDM will enable richer searching and browsing. This will likely lead to modifications in the overall Europeana system for supporting searching and browsing.

HathiTrust's support for search includes the use of both Solr for full-text searches and Z39.50 searches of their bibliographic records in OCLC and their ILS, Aleph. The ability of HathiTrust to support large-scale searching is remarkably robust and was developed in an iterative, scalable fashion (HathiTrust 2012). HathiTrust receives MARC records from many sources; the inconsistency of the quality and information makes robust searching of bibliographic records a challenge, but the team has methodically ensured that search and presentation of search results will meet the needs of HathiTrust users in successfully discovering information.

Both NSDL and NINES use Apache Nutch to support enhanced searching of content in their repositories (Apache Foundation 2011). Though NINES is a federated collection, it uses Nutch to crawl the content as well as the metadata so that there is a full-text index to support searches in NINES.

Search technologies continue to evolve and improve. The types of discovery services that digital libraries provide, the varying content formats, and overall performance considerations will shape decisions regarding the discovery technologies that should be included in a digital library platform.

5. Services and Applications

Large digital libraries are focusing more attention on modular development around services. A true service-oriented architecture (SOA) approach supports scalability and the addition, subtraction, or substitution of technologies over time. HathiTrust and NSDL are two large digital libraries that have embraced this approach in their system development efforts. Defining the functional components of the digital library in terms of services allows changes to be made in isolated or semi-isolated parts of the code with little impact on the other software components. It is easier to integrate new capabilities and improved technologies into the system if designers follow SOA principles while developing the system. Once defined and developed, services can be reused or modified, supporting flexibility and a modular architecture.

Efforts to define services specifically for education and libraries have begun, but there is no complete registry for all services that a digital library needs. Reusing services already defined for other SOA projects, such as the planned services for Project Bamboo, will enable flexibility and sustainability (HathiTrust 2012). The JISC e-Framework Initiative maintains a registry of reusable services that would likely be of use and would provide a development community for further open source development (eFramework Partners 2010). In 2006, the Digital Library Federation (DLF) explored the establishment of a services framework for digital libraries that could be used

for developing more flexible digital library systems (Lavoie, Henry, and Dempsey 2006). The microservices being developed by the CDL are used as finer-grained building blocks for modular development (Abrams, Kunze, and Loy 2010). All of these efforts can inform the development approaches for a large digital library that can be modified as new or improved functionality becomes available.

In addition to an overall SOA architecture, large digital libraries often provide value-added applications and services. NSDL provides a WordPress MultiUser blog, a MediaWiki, and Shibboleth user authentication. HathiTrust facilitates access to its content through a page-turner application and offers a Collection Builder interface. Europeana provides map and timeline views of its resources. NINES applications and services include

- Juxta, a bibliographic collation system
- IVANHOE, a multiplayer game of literary interpretation
- Collex, a tool for collecting and annotating digital objects and for publishing interlinked online exhibits that includes support for folksonomic tagging
- XML-to-RDF style sheets for TEI-encoded documents

Streaming services for audio and video can be included if the digital libraries support large audio and video files. JPEG2000 viewers can enhance image viewing for digital libraries with image collections. Support for specialized formats such as CAD-CAM architecture drawings can also be added.

These are just a few of the applications and services being provided by a handful of large digital libraries. They provide many more that help meet their users' needs in working with digital resources. With the proper systems architecture, these libraries can add new services as needed to ensure that they are providing value to their users.

6. System Sustainability

Although the infrastructure of the digital library is important for providing a robust management system for digital content, it is equally important that the system continue to operate reliably into the future. Thinking about sustainability—not just of the hardware and software, but also of the entire organization—early in the development process can help create a successful digital library that users can trust.

A realistic assessment to determine whether there is sufficient commitment to the level of staffing needed to keep the system functioning at all levels may drive the decision to have a federated collection or to manage the content centrally. Sustainable systems must be easily maintained, and they need to scale easily to meet growing traffic and content. Overall sustainability can be a problem for federated systems because the content is beyond the control of the library that people rely on for discovery and enhanced services.

Sustainability can be a key factor in users' trust of the system.

As they retrieve and reference resources, users need to know that those resources will be available to them over time. Therefore, it is important for large digital libraries to demonstrate they are a viable, trustworthy resource by making their efforts clear and by publishing their policies. The *Framework of Guidance for Building Good Digital Collections* provides guidance for both data providers and service providers to ensure that shared collections are meaningful and reliable. "Adherence to appropriate standards and collaboration between data providers and service providers are crucial elements of effective aggregated digital collections" (National Information Standards Organization [NISO] Framework Working Group 2007).

It is not sufficient for a repository simply to claim that it can be trusted to preserve its digital content. The Trustworthy Repository Audit and Certification (TRAC) process guides developers in creating a fully trusted repository in which users can have confidence. The requirements for TRAC certification are defined in the NISO Recommended Practice document, *Audit and Certification of Trustworthy Digital Repositories* (The Consultative Committee for Space Data Systems [CCSD] 2011).

The Europeana digital library has been diligent about publishing its policies and practices for the shared collections to which it provides access. Given the breadth of cultural heritage resources available through Europeana, the diversity of languages supported, and the types of organizations that contribute data, clearly defined policies are important to ensure reliable and predictable performance. Data providers are instructed to provide metadata in specific formats and to include specific elements. Tools are provided to help content contributors map their data to the required Europeana formats. Contributors must provide a persistent identifier so that the resources can be reliably accessed from the Europeana metadata. Content cannot have visible watermarks, ensuring that the online experience will be visually appealing. As a service provider, Europeana enhances the contributor-provided metadata by including standard multilingual terms and references to answer who, what, where, and when questions. Personal names are given a unique identifier, supporting links that allow users to discover more information about that person. Use of the GEMET thesaurus (European Environment Agency 2011) supports unique references for concepts, their display in multiple languages, and the display of references and labels of more general terms associated with the concept (Joyce Ray, personal communication).

HathiTrust has given serious consideration to its long-term sustainability. At the system level, it has implemented a modular architecture where discrete functional services are integrated for effective and efficient operation. This structure supports quick resolution of specific issues and distributed software development. Open standards and open systems enable partners to develop new services and components for repository functionality, thereby leveraging the expertise and contributions of the larger community. HathiTrust is TRAC certified, demonstrating that it meets strict

standards for trustworthiness. It has in place good software development practices, such as the use of a concurrent versioning system (CVS) repository to support code versioning for software consistency. The system adheres to University of Michigan Library security policies. HathiTrust has identified its approach to providing long-term preservation and curatorial services for content. There is a plan in place for disaster recovery. All of these activities, along with other policies and practices, provide evidence of a thoughtful, sustainable digital library.

7. Summary

The decision to establish a large digital library leads necessarily to a complex set of considerations. Decisions in one area will affect decisions in other areas. The focus of this study has been on understanding the core infrastructure elements of a few large digital libraries with diverse approaches that can serve as models for a large-scale digital library for U.S. cultural heritage assets. There are many more elements that can be examined, but it is critical to address the ones identified in this report to ensure that a system will not need to be reworked after a significant investment has been made to launch it. Much can be learned from the successes and challenges of other digital libraries.

In sum, the following considerations should be kept in mind during the planning and development phases for a large digital library. Scalability is critical to support long-term growth of the system, so the architecture decisions should support this. Building a modular system, following SOA principles, will enable flexibility, code reusability, and stronger system sustainability. It is essential to understand who the target audience is and what their needs are when interacting with the digital library. Talking with users and documenting the ways in which they would use the system are important to ensure that the system is appropriate. It is also important to decide on a realistic sustainability plan and to publish the policies and guidelines that will help enforce that plan.

As implementation begins, good project planning will keep the project on track and within scope. Developing documented usage scenarios to guide decisions about the most important functions can help. The most important functions, especially those that affect the overall architecture, need to be given priority over the "nice-to-have" features that can be implemented after the core system has stabilized. Although decisions about infrastructure can be based on research and discussion alone, a better approach will be to establish a sandbox environment to experiment with differing technologies and architectures. This approach will also help with launching beta code in the open source "release early, release often" ethos.

Large digital libraries that want to promote community involvement throughout the system development process and ongoing contribution of content will need strong communication approaches that keep the community involved and informed. With proper

management, communication will invite participation from those whose expertise can contribute to the common good that the large digital library strives to provide.

Attention to the issues discussed in this paper will lay a strong foundation for the complex endeavor of building a large digital library that can be sustained over time. Proceeding without early attention to these elements will put any digital library project at risk of failing or requiring a costly redesign down the road, especially if the system is intended to support very large amounts of content. Planning and experimentation early in the process will be key in guiding the project to a successful implementation.

References

Abrams, S., J. Kunze, and D. Loy. 2010. An Emergent Micro-services Approach to Digital Curation Infrastructure. *International Journal of Digital Curation* 5(1).

Amazon Web Services, LLC. n.d. Amazon Simple Storage Service (Amazon S3). Available at http://aws.amazon.com/s3/#pricing.

Apache Foundation. 2011. FrontPage: Apache Nutch Wiki. *Nutch* (November 27). Available at http://wiki.apache.org/nutch/.

California Digital Library. 2011. CDL's DPLA Vertical Search Demo. *DPLA Vertical Search Demo*. Available at http://crawlspace.cdlib.org/.

Center for Informatics Research in Science and Scholarship. Opening History: U.S. History Resources from Libraries, Museums, and Archives. 2009. *IMLS Digital Collections and Content*. Available at http://imlsdcc.grainger.uiuc.edu/history/.

Cornell University. 2006. MPTStore 0.9.1 Documentation. *MPTStore 0.9.1*. Available at http://mptstore.sourceforge.net/.

Coyle, Karen. 2006. Mass Digitization of Books. *The Journal of Academic Librarianship* 32(6): 641–645.

DataDirect Networks, Inc. 2012. Scalable Storage | Big Data Storage | DDN | DataDirect Networks™. *DataDirect Networks: Information in Motion*. Available at http://www.ddn.com/.

Dekkers, Makx, Stefan Gradmann, Carlo Meghini, Catherine Lupovici, Go Sugimoto, Robina Clayphan, Julie Verleyen, et al. 2011. *Europeana Semantic Elements Specifications* Version 3.4 (March 31).

Dekkers, Makx, Stefan Gradmann, and Jan Molendijk. 2011. *D3.4 Final Technical & Logical Architecture and Future Work Recommendations. Europeana Group* (October 5). Available at http://www.version1.europeana.eu/c/document_library/get_file?uuid=d0327b50-2e86-45bd-81c1-7bff4b9a449b&groupId=10602.

Digital Commonwealth. 2007. Digital Commonwealth. *Digital Commonwealth: Massachusetts Collections Online*. Available at http://www.digitalcommonwealth.org/.

eFramework Partners. 2010. e-Framework for Education and Research. *Service Genre Registry* (April 9). Available at http://www.e-framework.org/Default.aspx?tabid=987.

European Environment Agency. 2011. EIONET GEMET Thesaurus. *EEA EnviroWindows* (September 12). Available at http://www.eionet.europa.eu/gemet.

HathiTrust. 2012. Large-Scale Search | Www.hathitrust.org. *HathiTrust Digital Library*. Available at http://www.hathitrust.org/large_scale_search.

_____. n.d. Technological Profile | Www.hathitrust.org. *HathiTrust Digital Library: Technological Profile*. Available at http://www.hathitrust.org/technology.

Heath, Tom, and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space*. 1st ed. Synthesis Lectures on the Semantic Web: Theory and Technology 1:1 1–136. Morgan & Claypool Publishers. Available at http://linkeddatabook.com/editions/1.0/.

Isilon Systems. 2011. OneFS Operating System | Isilon Systems. *Isilon Systems*. Available at http://www.isilon.com/onefs-operating-system.

Krafft, D. B., A. Birkland, and E. J. Cramer. 2008. Ncore: Architecture and Implementation of a Flexible, Collaborative Digital Library. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 313–322. New York: ACM Press. Available at http://arxiv.org/abs/0803.1500v1.

Lavoie, Brian, Geneva Henry, and Lorcan Dempsey. 2006. A Service Framework for Libraries. *D-Lib Magazine* 12(7/8). Available at http://www.dlib.org/dlib/july06/lavoie/07lavoie.html.

National Information Standards Organization (NISO) Framework Working Group. 2007. *A Framework of Guidance for Building Good Digital Collections,* 3rd ed. A NISO Recommended Practice. Baltimore, MD: National Information Standards Organization (NISO). Available at http://www.niso.org/publications/rp/framework3.pdf.

nines.org. 2012. N I N E S. *NINES: Nineteenth-century Scholarship Online*. Available at http://www.nines.org/.

Nowviskie, Bethany, and Jerome McGann. 2005. NINES: A Federated Model for Integrating Digital Scholarship. Available at http://www.nines.org/about/wp-content/uploads/2011/12/9swhitepaper.pdf.

SeekingMichigan.org. 2008–2012. Seeking Michigan. Available at http://seekingmichigan.org/.

The Apache Software Foundation. n.d. Welcome to Apache Lucene! Available at http://lucene.apache.org/.

_____. n.d. Welcome to Solr. Available at http://lucene.apache.org/solr/.

The Consultative Committee for Space Data Systems (CCSD). 2011. *Audit and Certification of Trustworthy Digital Repositories, Issue 1*. Recommended Practice. Washington, D.C.: CCSDS Secretariat. Available at http://public.ccsds.org/publications/archive/652x0m1.pdf.

Additional Resources

Digital Library of Georgia. 2012. About the Digital Library of Georgia. *Digital Library of Georgia: Sharing Georgia's History and Culture Online* (January 27). Available at http://dlg.galileo.usg.edu/AboutDLG/.

Project Bamboo. 2012. Project Bamboo Technology Wiki - Home. Wiki. *Project Bamboo Technology Wiki* (February 1). Available at https://wiki.projectbamboo.org/display/BTECH/Technology+Wiki+-+Home.

Schmitz, Dawn. 2008. The Seamless Cyberinfrastructure: The Challenges of Studying Users of Mass Digitization and Institutional Repositories. Washington, DC: Council on Library and Information Resources. Available at www.clir.org/pubs/archives/schmitz.pdf.

State Library and Archives of Florida. 2012. Florida Memory. Available at http://www.floridamemory.com/.

Vancis BV. n.d. Vancis BV—A Subsidiary of SARA | Vancis. *Vancis Advanced ICT Services*. Available at http://www.vancis.nl/en/.

W3C. 2010. RDF—Semantic Web Standards. *W3C Semantic Web* (March 7). Available at http://www.w3.org/RDF/.

Wikipedia. n.d. Apache Solr. *Wikipedia, the Free Encyclopedia.* Available at http://en.wikipedia.org/wiki/Apache_Solr.

_____. n.d. Lucene. *Wikipedia, the Free Encyclopedia*. Available at http://en.wikipedia.org/wiki/Lucene.